First Hit Fwd Refs Previous Doc Next Doc Go to Doc#

End of Result Set

☑ Generate Collection Print

L5: Entry 1 of 1 File: USPT Dec 21, 1999

DOCUMENT-IDENTIFIER: US 6006221 A
TITLE: Multilingual document retrieval system and method using semantic vector matching

INVENTOR (4):
Li; Ming

Brief Summary Text (11):
The query's representation is then compared to each document's representation to generate a measure of relevance of the document to the query. Results can be browsed using a graphical interface, and individual documents (or document clusters) that seem highly relevant can be used to inform subsequent queries for relevance feedback. The system may also perform a surface-level, gloss transliteration of the foreign text, sufficient enough for a non-fluent reader to gain a basic understanding of the document's contents.

Detailed Description Text (55):
multilingual concept group n-gram probability database,

Detailed Description Text (72):
LI 120 determines by means of a combination of n-gram and word frequency analysis the language of the input document. The output of the LI is the document plus its language identification tag.

Detailed Description Text (73):
Two parallel approaches for language identification are employed. The first approach operates by scanning documents for a distribution of language-discriminant, common single words. The occurrence, frequency and distribution of these words in a document is compared against the same distributions gathered from a representative corpus of documents in each of the supported languages. The second approach involves locating common word/character sequences unique to each language. Such sequences may form actual words that often occur, such as conjunctions, or a mix of words, punctuation and character strings. Language identification involves scanning each document until a target character sequence is located.

Detailed Description Text (93):
The first of these modules, MCGRE 150, accepts the language-identified, part-of-speech tagged, input text and retrieves from the multilingual concept database any and all of the concept groups to which each input word belongs. Polysemous words (those words with multiple meanings) will have multiple concept group assignments at this stage. The output of the MCGRE 150, when run over a document, will be sentence-delimited strings of words, each word or phrase of which has been tagged with the codes of all the multilingual concept groups to which various senses of the word/phrase belongs.

Detailed Description Text (120):
The multilingual concept group n-gram probability database is an optional knowledge database that is constructed from a training data set. The database contents are

derived from a text corpus analysis of words used in various supported languages in various contexts. The data in the database can be either (1) sense-correct concept groups assigned to each term in the texts, or (2) all possible concept groups assigned to each term in the texts (e.g., if one term belongs to three concept groups, then three concept groups will be assigned to that term).

Detailed Description Text (122):
FIG. 3B shows this process where MCGD 160 has had to resort to Domain Knowledge (using the MCGCM) and Global Knowledge (using the n-gram probability database) to disambiguate the polysemous words.

Detailed Description Text (144):
(b) The resulting normalized document vectors are fixed-dimension vectors representing the concept-level contents of the processed text (either documents or queries). They are passed to the next module for either document-to-query-vector matching (comparison), or for document-to-document matching (comparison) for clustering of documents.

Detailed Description Text (185):
The fact that the documents are represented in a common, language-independent vector format of weighted slot values, no matter what the language of the individual documents, enables the system to treat all documents similarly. Therefore, it can: (1) cluster documents based on similarity amongst them, and (2) provide a single list of documents ranked by relevancy, with documents of various languages interfiled. Thus the process whereby documents are retrieved and ranked for review by the user is language independent.

Detailed Description Text (208):
GUI 250 uses clustering techniques to display conceptually-similar documents. The GUI also allows users to interact with the system by invoking relevance feedback, whereby a selection of documents or a single document can be used as the basis for a reformulated query to find those documents with conceptually similar contents.

Detailed Description Text (210):
8.3 Document Clustering, Browsing and Relevance Feedback

Detailed Description Text (211):
The monolingual category vectors are used as the basis for the clustering and display, and for the implementation of relevance feedback in the system:

Detailed Description Text (212):
8.3.1 Clustering

Detailed Description Text (213):
Documents can be clustered using an agglomerative (hierarchical) algorithm that compares all document vectors and creates clusters of documents with similarly weighted vectors. The nearest neighbor/Ward's approach is used to determine clusters, thus not forcing uniform sized clusters, and allowing new clusters to emerge when documents reflecting new subject areas are added. These agglomerative techniques, or divisive techniques, are appropriate because they do not require the imposition of a fixed number of clusters.

Detailed Description Text (214):
Using the clustering algorithm described above, or other algorithms such as single link or nearest neighbor, CINDOR is capable of mining large data sets and extracting highly relevant documents arranged as conceptually-related clusters in which documents from several languages co-occur.

Detailed Description Text (215):
Headlines from newspaper articles or titles from documents in the cluster are used

to form labels for clusters. Headlines or titles are selected from documents that are near the centroid of a particular cluster, and are therefore highly representative of the cluster's document contents. An alternative labeling scheme, selectable by the user, is the use of the labeled subject codes which make up either the centroid document's vector or the cluster vector.

Detailed Description Text (216):
The user is able to browse the documents, freely moving from cluster to cluster with the ability to view the full documents in addition to their summary representation. The user is able to indicate those documents deemed most relevant by highlighting document titles or summaries. If the user so decides, the relevance feedback steps can be implemented and an "informed" query can be produced, as discussed below.

Detailed Description Text (217):
The CINDOR system is thus able to display a series of conceptually-related clusters in response to a browsing query. Each cluster, or a series of clusters, could be used as a point of departure for further browsing. Documents indicative of a cluster's thematic and conceptual content would be used to generate future queries, thereby incorporating relevance feedback into the browsing process. The facility for browsing smaller, semantically similar sub-collections which contain documents of multiple languages aids users in determining which documents they might choose to have translated.

Detailed Description Text (219):
Relevance feedback is accomplished by combining the vectors of user-selected documents or document clusters with the original query vector to produce a new, "informed" query vector. The "informed" query vector will be matched against all document vectors in the corpus or those that have already passed the cut-off filter. Relevant documents will be re-ranked and re-clustered.

Detailed Description Text (220):
1. Combining of Vectors. The vector for the original query and all user-selected documents are weighted and combined to form a new, single vector for re-ranking and re-Clustering.

Detailed Description Text (222):
3. Cut-Off and Clustering after Relevance Feedback. Using the same regression formula described above in connection with recall predictor 240, a revised similarity score cut-off criterion is determined by the system on the basis of the "informed" query. The regression criteria are the same as for the original query, except that only the vector similarity score is considered. The agglomerative (hierarchical) clustering algorithm is applied to the vectors of the documents above the revised cut-off criterion and a re-clustering of the documents will be performed. Given the re-application of the cut-off criterion, the number of document vectors being clustered will be reduced, and improved clustering is achieved.

Detailed Description Text (228):
Documents or document clusters that, based on their high relevance ranking, the gloss transliteration, or other factors, are deemed to be highly relevant to a query, and are candidates for a machine translation of the original foreign language text. CINDOR thus ensures that only those few documents that are especially pertinent to a query will undergo the full translation process.

Detailed Description Text (239):
Another language-independent method of representing text is using n-gram coding, wherein a text is decomposed to a sequence of character strings, where each string contains n adjacent characters from the text. This can be done by moving an n-character window n characters at a time, or by moving the n-character window one

character at a time. In an n-gram representation, no attempt is made to understand, interpret or otherwise catalog the meaning of the text, or the words that make up the text. A tri-gram representation is the special case where n=3. Representation and matching are based on the co-occurrence of n-grams or a sequence of character strings, or on the co-occurrence and relative prevalence of such n-grams, or on other, similar schemes. Such analysis is an alternative representational scheme for CINDOR.

Detailed Description Text (240):

In this alternative embodiment, an n-gram query processor (NQP) module replaces probabilistic query processor (PQP) 220, an n-gram document processor replaces probabilistic term Indexer (PTI) 210, and an n-gram query to document matcher replaces query to document matcher (QDM 230). The NQP accepts the native-language input and performs the following processing: a) decomposes each term in the queries into n-adjacent-character strings; and b) lists each unique n-adjacent-character string with the number of occurrences as the document representation. The NDP accepts the output from PNC 140 and performs the following processing: a) decomposes each term in the document into n-adjacent-character strings; and b) lists each unique n-adjacent-character string with the number of occurrences as the query representation. The NQDM accepts two input streams, namely the outputs from the NQP and NDP, and provides a score representing the match between the documents and query. This output is an input to the score combiner. Documents are assigned scores by measuring the degrees of overlap between the n-gram decomposed terms from documents and queries. The larger the overlap, the higher the degree of relevance.

Current US Original Classification (1):

707/5

CLAIMS:

11. The method of claim 1, and further comprising:

determining a measure of proximity of the language-independent conceptual representation of each document to the language-independent conceptual representation of the other documents in the plurality; and

clustering the documents in the plurality according to the documents' respective measures of proximity to each other.

21. The method of claim 12 wherein said language-independent conceptual representation of the subject content of the document includes a statistical index using N-gram style decomposed words as indexing units.

23. The method of claim 12 wherein said language-independent conceptual representation of the subject content of the query includes N-gram style decomposed terms as language-independent query requirements.

32. The method of claim 12 wherein generating a measure of relevance for a given document comprises:

generating an N-gram decomposed term representation for the given document and for the query; and

determining a degree of overlap between the N-gram decomposed terms, the overlap representing the measure of relevance, with a larger overlap representing a higher degree of relevance.

Previous Doc Next Doc Go to Doc#

First Hit Fwd Refs Previous Doc Next Doc Go to Doc#

End of Result Set

□ Generate Collection Print

L6: Entry 1 of 1 File: USPT Dec 21, 1999

DOCUMENT-IDENTIFIER: US 6006221 A
TITLE: Multilingual document retrieval system and method using semantic vector matching

INVENTOR (4):
Li; Ming

Brief Summary Text (11):
The query's representation is then compared to each document's representation to generate a measure of relevance of the document to the query. Results can be browsed using a graphical interface, and individual documents (or document clusters) that seem highly relevant can be used to inform subsequent queries for relevance feedback. The system may also perform a surface-level, gloss transliteration of the foreign text, sufficient enough for a non-fluent reader to gain a basic understanding of the document's contents.

Detailed Description Text (25):
Processing of documents and queries follows a modular progression, with documents being matched to queries based on matching (1) their conceptual-level contents, and (2) various term-based and logic representations such as the frequency/co-occurrence of proper nouns. At the conceptual level of matching, each substantive word in a document or query is assigned a concept category, and these category frequencies are summed to produce a vector representation of the whole text. Proper nouns are considered separately and, using a modified, fuzzy Boolean representation, matching occurs based on the frequency and co-occurrence of proper nouns in documents and queries. The principles applied to the proper noun matching are applicable to matching for other terms and parts of speech, such as complex nominals (CNs) and single terms.

Detailed Description Text (55):
multilingual concept group n-gram probability database,

Detailed Description Text (57):
frequency database.

Detailed Description Text (72):
LI 120 determines by means of a combination of n-gram and word frequency analysis the language of the input document. The output of the LI is the document plus its language identification tag.

Detailed Description Text (73):
Two parallel approaches for language identification are employed. The first approach operates by scanning documents for a distribution of language-discriminant, common single words. The occurrence, frequency and distribution of these words in a document is compared against the same distributions gathered from a representative corpus of documents in each of the supported languages. The second approach involves locating common word/character sequences unique to each language. Such sequences may form actual words that often occur, such as conjunctions, or a

mix of words, punctuation and character strings. Language identification involves scanning each document until a target character sequence is located.

Detailed Description Text (93):
The first of these modules, MCGRE 150, accepts the language-identified, part-of-speech tagged, input text and retrieves from the multilingual concept database any and all of the concept groups to which each input word belongs. Polysemous words (those words with multiple meanings) will have multiple concept group assignments at this stage. The output of the MCGRE 150, when run over a document, will be sentence-delimited strings of words, each word or phrase of which has been tagged with the codes of all the multilingual concept groups to which various senses of the word/phrase belongs.

Detailed Description Text (115):
Global Knowledge simulates the observation made in human sense disambiguation that more frequently used senses of words are cognitively activated in preference to less frequently used senses of words. Therefore, the words not yet disambiguated by Local Context or Domain Knowledge will now have their multiple concept group codes compared to a Global Knowledge database source, referred to as the frequency database. The database is an external, off-line sense-tagging of parallel corpora with the correct concept group code for each word. The disambiguated parallel corpora will provide frequencies of each word's usage as a particular sense (equatable to concept group) in the sample corpora. The most frequent sense is selected as the concept category.

Detailed Description Text (116):
The frequency database can be constructed in any of the following three ways:

Detailed Description Text (117):
(1) Collect the most frequent sense information from partially or fully sense-disambiguated texts (the training data to collect sense frequency information can be built either manually or automatically). Training data can be built automatically from the output from MCGD module without the frequency database OR the output from automatic sense comparison using multilingual aligned corpus such as "Canadian Hansard."

Detailed Description Text (119):
(3) Use frequency information from a lexicon that provides its senses with frequency information.

Detailed Description Text (120):
The multilingual concept group n-gram probability database is an optional knowledge database that is constructed from a training data set. The database contents are derived from a text corpus analysis of words used in various supported languages in various contexts. The data in the database can be either (1) sense-correct concept groups assigned to each term in the texts, or (2) all possible concept groups assigned to each term in the texts (e.g., if one term belongs to three concept groups, then three concept groups will be assigned to that term).

Detailed Description Text (122):
FIG. 3B shows this process where MCGD 160 has had to resort to Domain Knowledge (using the MCGCM) and Global Knowledge (using the n-gram probability database) to disambiguate the polysemous words.

Detailed Description Text (127):
(b) Converts the English word members of the selected concept group from the multilingual concept database (MCD) to zero or more categories in the monolingual hierarchical concept dictionary (MHCD). This is a static mapping scheme, whereby all the English word members of a particular concept group are treated as being equally likely instantiations. In this static implementation, all English word

members of the selected multilingual concept group are mapped to their respective categories in the MHCD. The frequencies of the concept categories mapped to by the English word members of the selected multilingual concept group of a word are summed and the most frequent category for that word is selected. If there are multiple categories in the MHCD to which the English word members of the multilingual concept group map, then these multiple categories need to be disambiguated in the next component of the system.

Detailed Description Text (136):
(c) Global Knowledge--If there is no Unique or Frequent monolingual category in an input sentence, then the system has no "anchor" by which to access the Correlation Matrix and must use global knowledge. In this event, the frequency of use of various senses of a word is used as the basis for the global knowledge source.

Detailed Description Text (140):
The MCVG generates a representation of the meaning (context) of the text of a document/query in the form of monolingual category (subject) codes assigned to information bearing words in the text. The monolingual category vector for all documents and queries has the same number of dimensions; weights or scores are applied to each dimension according to the presence and frequency of text with certain subject-contents.

Detailed Description Text (143):
(a) The frequencies of the disambiguated monolingual category codes assigned to words in the text are summed and then normalized in order to control for the effect of document length.

Detailed Description Text (144):
(b) The resulting normalized document vectors are fixed-dimension vectors representing the concept-level contents of the processed text (either documents or queries). They are passed to the next module for either document-to-query-vector matching (comparison), or for document-to-document matching (comparison) for clustering of documents.

Detailed Description Text (153):
PTI 210 accepts the output from PNC 140 (documents only) and creates a new appended field in the document index file. The PTI also assigns a weighted, TF.IDF score (the product of Term Frequency and Inverse Document Frequency) for each proper noun. This could be applied to other types of terms. This weighted score is used in QDM and score combiner 230. This index file contains all proper nouns and their associated TF.IDF scores.

Detailed Description Text (155):
where TF is the number of occurrences of a term within a given document, IDF is the inverse of the number of documents in which the term occurs, compared to the whole corpus, N is the total number of documents in the corpus, and n is the number of documents in which the term occurs. The product of TF.IDF provides a quantitative indication of a term's relative uniqueness and importance for matching purposes. TF.IDF scores are calculated for documents and queries. The IDF scores are based upon the frequency of occurrence of terms within a large, representative sample of documents in each supported language.

Detailed Description Text (162):
For each supported language there exists a class of frequently used words or phrases that, when connected in a logical sequence, are used to establish the transition from the positive to the negative portion of the query (or the reverse). In English such a sequence might be as simple as "I am interested in" followed by ", but not." Clue words or phrases must have a high frequency of occurrence within the confines of a particular context.

Detailed Description Text (184):
Documents are arranged in ranked order according to their relative relevance to the substance of a query. The matcher uses a variety of evidence sources to determine the similarity or suitable association between query and documents. Various representations of document and query are used for matching, and each document-query pair is assigned a match score based on (1) the distance between vectors, and (2) the frequency and occurrence of proper nouns.

Detailed Description Text (185):
The fact that the documents are represented in a common, language-independent vector format of weighted slot values, no matter what the language of the individual documents, enables the system to treat all documents similarly. Therefore, it can: (1) cluster documents based on similarity amongst them, and (2) provide a single list of documents ranked by relevancy, with documents of various languages interfiled. Thus the process whereby documents are retrieved and ranked for review by the user is language independent.

Detailed Description Text (208):
GUI 250 uses clustering techniques to display conceptually-similar documents. The GUI also allows users to interact with the system by invoking relevance feedback, whereby a selection of documents or a single document can be used as the basis for a reformulated query to find those documents with conceptually similar contents.

Detailed Description Text (210):
8.3 Document Clustering, Browsing and Relevance Feedback

Detailed Description Text (211):
The monolingual category vectors are used as the basis for the clustering and display, and for the implementation of relevance feedback in the system:

Detailed Description Text (212):
8.3.1 Clustering

Detailed Description Text (213):
Documents can be clustered using an agglomerative (hierarchical) algorithm that compares all document vectors and creates clusters of documents with similarly weighted vectors. The nearest neighbor/Ward's approach is used to determine clusters, thus not forcing uniform sized clusters, and allowing new clusters to emerge when documents reflecting new subject areas are added. These agglomerative techniques, or divisive techniques, are appropriate because they do not require the imposition of a fixed number of clusters.

Detailed Description Text (214):
Using the clustering algorithm described above, or other algorithms such as single link or nearest neighbor, CINDOR is capable of mining large data sets and extracting highly relevant documents arranged as conceptually-related clusters in which documents from several languages co-occur.

Detailed Description Text (215):
Headlines from newspaper articles or titles from documents in the cluster are used to form labels for clusters. Headlines or titles are selected from documents that are near the centroid of a particular cluster, and are therefore highly representative of the cluster's document contents. An alternative labeling scheme, selectable by the user, is the use of the labeled subject codes which make up either the centroid document's vector or the cluster vector.

Detailed Description Text (216):
The user is able to browse the documents, freely moving from cluster to cluster with the ability to view the full documents in addition to their summary representation. The user is able to indicate those documents deemed most relevant

by highlighting document titles or summaries. If the user so decides, the relevance feedback steps can be implemented and an "informed" query can be produced, as discussed below.

Detailed Description Text (217):
The CINDOR system is thus able to display a series of conceptually-related clusters in response to a browsing query. Each cluster, or a series of clusters, could be used as a point of departure for further browsing. Documents indicative of a cluster's thematic and conceptual content would be used to generate future queries, thereby incorporating relevance feedback into the browsing process. The facility for browsing smaller, semantically similar sub-collections which contain documents of multiple languages aids users in determining which documents they might choose to have translated.

Detailed Description Text (219):
Relevance feedback is accomplished by combining the vectors of user-selected documents or document clusters with the original query vector to produce a new, "informed" query vector. The "informed" query vector will be matched against all document vectors in the corpus or those that have already passed the cut-off filter. Relevant documents will be re-ranked and re-clustered.

Detailed Description Text (220):
1. Combining of Vectors. The vector for the original query and all user-selected documents are weighted and combined to form a new, single vector for re-ranking and re-Clustering.

Detailed Description Text (222):
3. Cut-Off and Clustering after Relevance Feedback. Using the same regression formula described above in connection with recall predictor 240, a revised similarity score cut-off criterion is determined by the system on the basis of the "informed" query. The regression criteria are the same as for the original query, except that only the vector similarity score is considered. The agglomerative (hierarchical) clustering algorithm is applied to the vectors of the documents above the revised cut-off criterion and a re-clustering of the documents will be performed. Given the re-application of the cut-off criterion, the number of document vectors being clustered will be reduced, and improved clustering is achieved.

Detailed Description Text (228):
Documents or document clusters that, based on their high relevance ranking, the gloss transliteration, or other factors, are deemed to be highly relevant to a query, and are candidates for a machine translation of the original foreign language text. CINDOR thus ensures that only those few documents that are especially pertinent to a query will undergo the full translation process.

Detailed Description Text (239):
Another language-independent method of representing text is using n-gram coding, wherein a text is decomposed to a sequence of character strings, where each string contains n adjacent characters from the text. This can be done by moving an n-character window n characters at a time, or by moving the n-character window one character at a time. In an n-gram representation, no attempt is made to understand, interpret or otherwise catalog the meaning of the text, or the words that make up the text. A tri-gram representation is the special case where n=3. Representation and matching are based on the co-occurrence of n-grams or a sequence of character strings, or on the co-occurrence and relative prevalence of such n-grams, or on other, similar schemes. Such analysis is an alternative representational scheme for CINDOR.

Detailed Description Text (240):
In this alternative embodiment, an n-gram query processor (NQP) module replaces

probabilistic query processor (PQP) 220, an n-gram document processor replaces probabilistic term Indexer (PTI) 210, and an n-gram query to document matcher replaces query to document matcher (QDM 230). The NQP accepts the native-language input and performs the following processing: a) decomposes each term in the queries into n-adjacent-character strings; and b) lists each unique n-adjacent-character string with the number of occurrences as the document representation. The NDP accepts the output from PNC 140 and performs the following processing: a) decomposes each term in the document into n-adjacent-character strings; and b) lists each unique n-adjacent-character string with the number of occurrences as the query representation. The NQDM accepts two input streams, namely the outputs from the NQP and NDP, and provides a score representing the match between the documents and query. This output is an input to the score combiner. Documents are assigned scores by measuring the degrees of overlap between the n-gram decomposed terms from documents and queries. The larger the overlap, the higher the degree of relevance.

Current US Original Classification (1):
707/5

CLAIMS:

11. The method of claim 1, and further comprising:

determining a measure of proximity of the language-independent conceptual representation of each document to the language-independent conceptual representation of the other documents in the plurality; and

clustering the documents in the plurality according to the documents' respective measures of proximity to each other.

21. The method of claim 12 wherein said language-independent conceptual representation of the subject content of the document includes a statistical index using N-gram style decomposed words as indexing units.

23. The method of claim 12 wherein said language-independent conceptual representation of the subject content of the query includes N-gram style decomposed terms as language-independent query requirements.

32. The method of claim 12 wherein generating a measure of relevance for a given document comprises:

generating an N-gram decomposed term representation for the given document and for the query; and

determining a degree of overlap between the N-gram decomposed terms, the overlap representing the measure of relevance, with a larger overlap representing a higher degree of relevance.

Previous Doc Next Doc Go to Doc#